Dimensionality reduction using Principal Component Analysis (PCA)

Problem

We are given the following dataset with two features, x_1 and x_2, and we want to reduce the dimensionality from 2D to 1D using PCA.

S. No.	x_1	x_2
1	4	1
2	2	4
3	2	3
4	3	6
5	4	4
6	9	10
7	6	8
8	9	5
9	8	7
10	10	8

Solution

Step 1: Mean of Each Feature

The mean of x_1 and x_2 are computed as follows:

1. For x_1:

$$\mu_{x_1} = \frac{4+2+2+3+4+9+6+9+8+10}{10} = 5.7$$

2. For x_2:

$$\mu_{x_2} = \frac{1+4+3+6+4+10+8+5+7+8}{10} = 5.6$$

Step 2: Covariance Matrix Elements

We compute each element of the covariance matrix using the covariance formula:

$$\text{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y)$$

Variance of x_1

$$\text{Cov}(x_1, x_1) = \frac{1}{9}\sum_{i=1}^{n}(x_{1,i} - \mu_{x_1})^2$$

Substituting the values:

$$\text{Cov}(x_1, x_1) = \frac{1}{9}\left[(4-5.7)^2 + (2-5.7)^2 + (2-5.7)^2 + (3-5.7)^2 + (4-5.7)^2 + (9-5.7)^2 + (6-5.7)^2 + (9-5.7)^2\right.$$

$$+(8-5.7)^2+(10-5.7)^2\big]$$

$$=\frac{1}{9}\left[2.89+13.69+13.69+7.29+2.89+10.89+0.09+10.89+5.29+18.49\right]$$

$$=\frac{1}{9}\times 86.1=9.567$$

So, $\text{Cov}(x_1, x_1) = 9.567$.

Covariance between x_1 and x_2

$$\text{Cov}(x_1,x_2)=\frac{1}{9}\sum_{i=1}^{n}(x_{1,i}-\mu_{x_1})(x_{2,i}-\mu_{x_2})$$

Substituting the values:

$$\text{Cov}(x_1,x_2)=\frac{1}{9}\left[(4-5.7)(1-5.6)+(2-5.7)(4-5.6)+(2-5.7)(3-5.6)+(3-5.7)(6-5.6)+(4-5.7)(4-5.6)\right.$$

$$\left.+(9-5.7)(10-5.6)+(6-5.7)(8-5.6)+(9-5.7)(5-5.6)+(8-5.7)(7-5.6)+(10-5.7)(8-5.6)\right]$$

$$=\frac{1}{9}\left[(-1.7\times -4.6)+(-3.7\times -1.6)+(-3.7\times -2.6)+(-2.7\times 0.4)+(-1.7\times -1.6)+(3.3\times 4.4)\right.$$

$$\left.+(0.3\times 2.4)+(3.3\times -0.6)+(2.3\times 1.4)+(4.3\times 2.4)\right]$$

$$=\frac{1}{9}\times 51.78=5.753$$

So, $\text{Cov}(x_1, x_2) = 5.753$.

Variance of x_2

$$\text{Cov}(x_2,x_2)=\frac{1}{9}\sum_{i=1}^{n}(x_{2,i}-\mu_{x_2})^2$$

Substituting the values:

$$\text{Cov}(x_2,x_2)=\frac{1}{9}\big[(1-5.6)^2+(4-5.6)^2+(3-5.6)^2+(6-5.6)^2+(4-5.6)^2+(10-5.6)^2+(8-5.6)^2$$

$$+(5-5.6)^2+(7-5.6)^2+(8-5.6)^2\big]$$

$$=\frac{1}{9}\left[21.16+2.56+6.76+0.16+2.56+19.36+5.76+0.36+1.96+5.76\right]$$

$$=\frac{1}{9}\times 65.4=7.267$$

So, $\text{Cov}(x_2, x_2) = 7.267$.
Thus, the final covariance matrix for the original data X is:

$$\text{Cov}(X)=\begin{bmatrix}\text{Cov}\left(x_1,x_1\right) & \text{Cov}\left(x_1,x_2\right)\\ \text{Cov}\left(x_1,x_2\right) & \text{Cov}\left(x_2,x_2\right)\end{bmatrix}=\begin{bmatrix}9.567 & 5.753\\ 5.753 & 7.267\end{bmatrix}$$

Step 3: Compute Eigenvalues and Eigenvectors

To find the principal components, we compute the eigenvalues and eigenvectors of the covariance matrix. We solve the equation:

$$Av = \lambda v$$

where $A = \mathrm{Cov}(X)$.
The eigenvalues are:

$$\lambda_1 = 14.331, \quad \lambda_2 = 2.614$$

The eigenvectors are:

$$v_1 = \begin{bmatrix} 0.77 \\ 0.638 \end{bmatrix}, \quad v_2 = \begin{bmatrix} -0.638 \\ 0.77 \end{bmatrix}$$

The eigenvector corresponding to the largest eigenvalue, v_1, defines the direction of the first principal component.

Step 4: Project the Data onto the First Principal Component

We project the standardized data onto the first principal component by multiplying each standardized data point by the first eigenvector:
For example, for the first data point $(4, 1)$:

$$\text{Transformed value} = (4)(0.77) + (1)(0.638) = 3.08 + 0.638 = 3.718$$

Similarly, we project all data points. The transformed data is:

$$\text{Transformed Data} = \begin{bmatrix} (4)(0.77) + (1)(0.638) \\ (2)(0.77) + (4)(0.638) \\ (2)(0.77) + (3)(0.638) \\ (3)(0.77) + (6)(0.638) \\ (4)(0.77) + (4)(0.638) \\ (9)(0.77) + (10)(0.638) \\ (6)(0.77) + (8)(0.638) \\ (9)(0.77) + (5)(0.638) \\ (8)(0.77) + (7)(0.638) \\ (10)(0.77) + (8)(0.638) \end{bmatrix} = \begin{bmatrix} 3.718 \\ 4.091 \\ 3.454 \\ 6.137 \\ 5.632 \\ 13.309 \\ 9.723 \\ 10.121 \\ 10.626 \\ 12.804 \end{bmatrix}$$

Final 1D Transformed Data

After performing PCA, the reduced 1D dataset is:

S. No.	x'_1
1	3.718
2	4.091
3	3.454
4	6.137
5	5.632
6	13.309
7	9.723
8	10.121
9	10.626
10	12.804

Summary

The steps of PCA are:

1. Compute mean values for each column.
2. Compute the covariance matrix.
3. Find the eigenvalues and eigenvectors of the covariance matrix.
4. Project the data onto the first n principal components to reduce the dimensionality.

Thus, we have successfully reduced the dimensions of the data using PCA.